Biostatistics (2004), **5**, 4, *pp.* 515–529
doi: 10.1093/biostatistics/kxh005

Identifying multiple changepoints in heterogeneous binary data with an application to molecular genetics

PAUL S. ALBERT*, SALLY A. HUNSBERGER

Biometric Research Branch, National Cancer Institute, 6130 Executive Blvd, Room 8136, Bethesda, MD 20892, USA

NAN HU, PHILIP R. TAYLOR

Cancer Prevention Studies Branch, National Cancer Institute, 6116 Executive Blvd, Room 705, Bethesda, MD 20892, USA

SUMMARY

Identifying changepoints is an important problem in molecular genetics. Our motivating example is from cancer genetics where interest focuses on identifying areas of a chromosome with an increased likelihood of a tumor suppressor gene. Loss of heterozygosity (LOH) is a binary measure of allelic loss in which abrupt changes in LOH frequency along the chromosome may identify boundaries indicative of a region containing a tumor suppressor gene. Our interest was on testing for the presence of multiple changepoints in order to identify regions of increased LOH frequency. A complicating factor is the substantial heterogeneity in LOH frequency across patients, where some patients have a very high LOH frequency while others have a low frequency. We develop a procedure for identifying multiple changepoints in heterogeneous binary data. We propose both approximate and full maximum-likelihood approaches and compare these two approaches with a naive approach in which we ignore the heterogeneity in the binary data. The methodology is used to estimate the pattern in LOH frequency on chromosome 13 in esophageal cancer patients and to isolate an area of inflated LOH frequency on chromosome 13 which may contain a tumor suppressor gene. Using simulations, we show that our approach works well and that it is robust to departures from some key modeling assumptions.

Keywords: Change point detection; Correlated binary data; Heterogeneity; Loss of heterozygosity; Repeated binary data; Spectral correlation.

1. INTRODUCTION

Loss of heterozygosity (LOH), sometimes referred to as allelic loss, is a term which indicates that at a particular marker location only one of two different parental alleles present in normal DNA is present in tumor cell DNA (Gruice *et al.*, 1993; Ross, 1998). LOH is often summarized as a dichotomous outcome which is positive if there is allelic loss and negative if not. This binary measure of genetic instability is relatively quick, simple, reproducible, and can be determined from archival material. Thus, LOH is particularly useful in population-based studies in which many samples obtained from archival material

*To whom correspondence should be addressed.

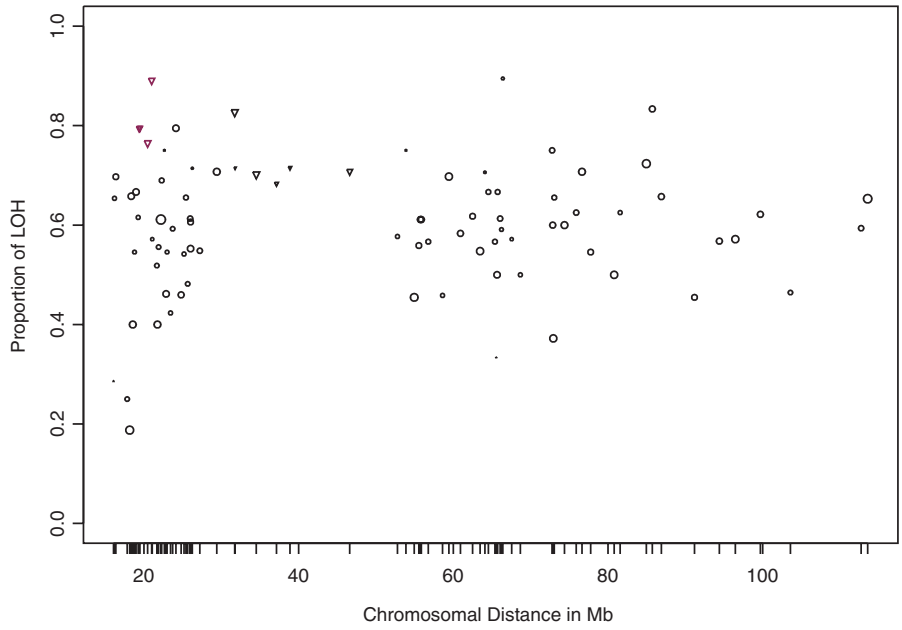

Fig. 1. Estimated LOH frequency for each of the 85 markers on chromosome 13. Area of the circle is proportional to the # of informative markers at that location; x-axis is in units of Mb from the p-telomere. The exact chromosomal location for the four markers between 17.85 and 18.95 Mb was unknown; however, their relative ordering is known. For simplicity, we placed the four markers equally spaced over the interval.

are analyzed. Fine mapping using numerous microsatellite markers is a standard approach for identifying chromosomal regions with high LOH frequency. Chromosomal regions with frequent allelic loss may point to tumor suppressor genes; identification of such genes may help in understanding the genetics of esophageal cancer, and lead to the development of early detection markers for this disease. Thus, we were interested in identifying boundaries which will define chromosomal regions with increased LOH frequency. The motivation for this work comes from a study examining LOH frequency in a group of 56 patients with esophageal cancer. LOH was measured at various markers on a series of chromosomes (Huang *et al.*, 2000; Hu *et al.*, 2000; Li *et al.*, 2001). The tumor marker is informative (i.e. one can measure LOH) at a given marker only if the normal DNA is heterozygous at that location. The proportion of informative markers varies widely by marker location. Figure 1 shows LOH frequency for 85 markers placed along chromosome 13. The x-axis for this plot is chromosomal distance in Mega bases (Mb) from the p-telomere (top of the short arm of the chromosome). The area of the plotting symbols in Figure 1 is related to the number of informative markers at a given marker location. The wide variation in the area of the points demonstrates that the proportion of informative markers varies widely by marker location. The proportion of informative markers ranged from 0.11 to 0.96 (median $= 0.57$) for chromosome 13. The identification of regions with particularly high LOH is not visually obvious from Figure 1. Thus, our goal was to develop a statistically sound approach for identifying regions with inflated LOH frequency.

A complicating factor in the analysis of these LOH data is the large between-subject variation in LOH frequency and the potential for marker measurements to be spatially correlated. Figure 2 shows a scatter plot of the log-odds ratio versus chromosomal distance (Mb) for chromosome 13. A LOWESS smoother (Cleveland, 1979) was applied to help identify the form of the dependence structure. The fact that the smoothed curve is almost a horizontal line suggests that there is very little spatial correlation

in these binary measurements. The nearly constant large log-odds ratio reflects very high exchangeable correlation (i.e. large between-subject variation). The large between-subject variation suggests that there is a large amount of heterogeneity in overall genetic instability across individuals. This heterogeneity needs to be accounted for in our analyses. A common strategy for identifying deletion regions is to examine the individual sequences along with the marginal frequencies. Such an 'eye-ball' approach may work well when deletion regions are obvious. Unfortunately, for esophageal cancer, where the overall LOH frequency is high (suggesting that the tumors are inherently genetically unstable), it is difficult to detect deletion regions with this approach. In fact, Li *et al.* (2001) reported that this overall high instability was striking, and has not been observed for other tumor types. For the esophageal cancer study, Li *et al.* (2001) used an ad hoc approach to identify deletion regions on chromosome 13. They defined regions by having at least 75% marginal LOH frequency over a contiguous region. Using this approach, they identified two inflated LOH regions over the first 42 markers ordered by distance from the p-telomere (at the time of their paper only data on the first 42 markers were available). Both regions are shown in Figure 1, where triangles represent LOH frequency within these regions and circles represent LOH frequency outside these regions. The first region includes four markers at locations 19.37, 19.49, 20.49, and 21.00 Mb (spanning a distance of 1.63 Mb), and has some important candidate genes such as ectodermal dysplasia 2, gap junction protein beta-2, and zinc finger protein 198, that may be tumor suppressor genes. The second region includes six markers located from 31.74 to 46.61 Mb (spanning a distance of 14.87 Mb), and contains two markers that are known to flank the BRCA2 gene which is under investigation for a potential role in esophageal cancer. Although interesting, the procedure for selecting these inflated regions is ad hoc and lacks statistical rigor. Our focus is on developing a more rigorous approach for the identification of deletion regions in this study. The strongest evidence for a deletion region is large abrupt changes in LOH frequency around that region. Thus, a multiple changepoint model formulation allows for such changes and easily allows for the estimation of boundaries surrounding deletion regions.

The identification of multiple changepoints in repeated binary data is a common problem in genetics, and this is an area in which various approaches have been developed. Much of this work focuses on trying to organize sequences of DNA base information into homogeneous segments; this work has been called segmentation analysis in the statistical genetics field. The typical data structure for this problem is a long sequence of binary data (on the order of 10 000 markers) that needs to be *segmented* into different regions with different mutation frequencies. Auger and Lawrence (1989) presented a computationally feasible algorithm for obtaining the maximum-likelihood for multiple changepoints in long sequences of independent data. Extensions of this work include Fu and Curnow (1990) who presented a maximum-likelihood approach for sequences of independent categorical data which allows for minimum-specified regions. Braun *et al.* (2000) developed methodology for fitting multiple changepoint quasi-likelihood models with applications to fitting repeated binomial data with over-dispersion. Hidden Markov chain models have been developed for estimating changepoints in long sequences (for example, Churchill, 1989). These models postulate that transitions between regions (i.e. changepoints) are governed by an unobserved Markov chain. Changepoint estimation involves estimating the probability of a transition given the data at each observation in the sequence. There have been a number of Bayesian approaches to the multiple changepoint problem (Carlin *et al.*, 1992; Stephens, 1994, among others). These models require the specification of a prior distribution on the number and position of the changepoints, and are computationally intensive.

In this paper we develop a frequentist approach for estimating multiple changepoint models for sequences of binary data with heterogeneity across subjects. Specifically, we focus on the situation where the binary sequences are long and are collected on multiple individuals as in our LOH/esophageal cancer example. Our work supplements the work of Newton and Lee (2000) who developed a stochastic modeling approach for inferring the location and effect of tumor suppressor genes. We present the changepoint modeling approaches in Section 2. We also propose a procedure for choosing the number of changepoints

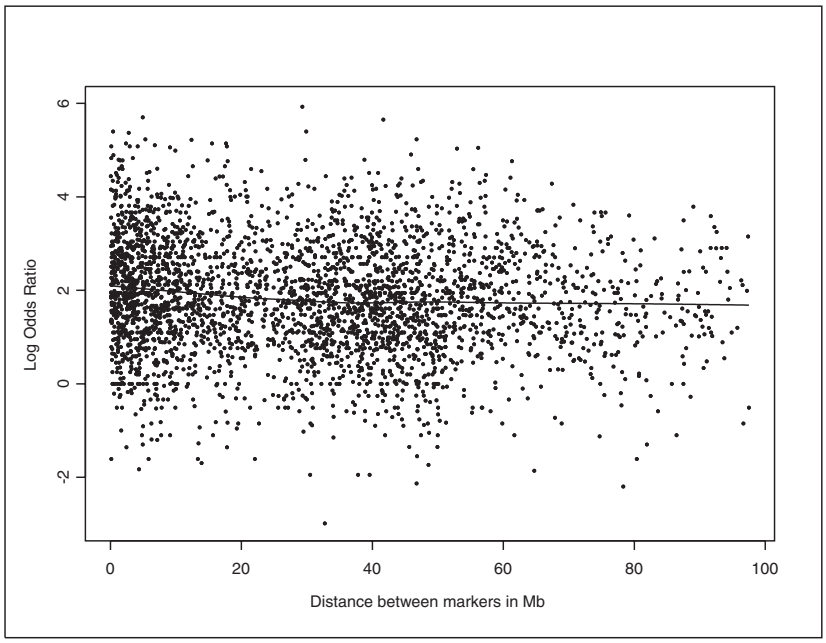

Fig. 2. Log-odds ratio versus distance in Mb between all pairwise combinations of markers. The smoothed line is a LOWESS smooth through the points.

which minimizes the chance of concluding that there are changepoint(s) when in fact there are none. We use these models to estimate the boundaries of regions with inflated LOH frequency on chromosome 13 in our esophageal cancer study in Section 3. In Section 4 we demonstrate the properties of our procedures with a series of simulation studies. The robustness of our approach to modeling assumptions is also examined. A discussion follows in Section 5.

2. MODEL FORMULATION

Let Y_{ij} be a binary response at each marker location for $j = 1, \ldots, J$ markers (ordered by distance from the p-telomere) and for each individual $i = 1, 2, \ldots, I$ where I and J are the total number of individuals and observations per individual, respectively. For R distinct regions with differing frequency, we assume that Y_{ij} is Bernoulli with proportion

$$\text{logit} P(Y_{ij} = 1) = \sum_{r=1}^{R} \beta_r I_{(\eta_{r-1}, \eta_r]} + b_i, \tag{1}$$

where b_i is a subject-specific intercept which reflects different overall LOH frequency between patients, β_r are the effects of being in the rth region for $r = 1, 2, \ldots, R$, and I_x is an indicator of being within the interval x. In addition, $\eta_0, \eta_1, \eta_2, \ldots, \eta_R$ are the boundaries of the R regions where η_0 and η_R reflect the beginning and end of the chromosome, respectively. Thus, we will estimate the $R - 1$ changepoints, $\eta_1, \eta_2, \ldots, \eta_{R-1}$, when we have R regions. Interest will focus on testing for the presence of disjoint changes in the mean structure as well as estimating the boundaries. Of particular importance, for our application, is to identify those regions with inflated LOH frequency. For simplicity we will assume that

the b_i are fixed effects. This is reasonable since we are considering the case in which we have many binary measurements on each individual (85 LOH markers on chromosome 13). Model (1) reduces to a Rasch model (Agresti, 1990) when $R = J$. Model (1) is also related to a model proposed by Korn and Whittemore (1979) for long sequences of binary data on multiple subjects where each individual has a unique fixed effect.

Note that model (1) only depends on chromosomal location in that markers need to be ordered by their location. This formulation does not depend on the actual chromosomal distances since the model assumes, conditional on b_i, that the sequences of binary data are spatially independent. This assumption is reasonable for our LOH example since Figure 2 demonstrates near spatial independence. Section 4 examines the effect of ignoring a moderate amount of spatial dependence on inference. Model (1) also assumes that frequency is constant within a region. We examine the robustness of our inferences to this assumption with a simulation presented in Section 4.

We wish to develop an approach for obtaining maximum-likelihood estimators of the model parameters and selecting the number of regions and estimating the boundaries. Direct maximization of the parameters for this model is computationally difficult for more then a few regions. Auger and Lawrence (1989) developed an approach for long sequences of independent continuous data. We extend this procedure to the situation where we have repeated binary data on multiple subjects.

We begin with the case of no patient heterogeneity (i.e. $b_i = 0$ in (1)) and no missing data. For a fixed number of regions R, the algorithm, which we will call the independence approach, is implemented with the following steps:

1. Evaluate the log-likelihood for regions $[l, m]$, $l \leqslant m$, with a constant frequency. We will denote each of the $J(J+1)/2$ log-likelihoods as
$$L^1_{l,m} = X_{lm}\log\widehat{p}_{lm} + (N_{lm} - X_{lm})\log(1 - \widehat{p}_{lm})$$
where p_{lm} is the probability of a positive response over the region $[l, m]$, X_{lm} and N_{lm} are the number of positive responses and the total number of binary markers for all I individuals over the region $[l, m]$, and $\widehat{p}_{lm} = X_{lm}/N_{lm}$. The superscript in $L^1_{l,m}$ reflects the number of regions with constant frequency (i.e. one region) over the interval $[l, m]$.
2. Construct the maximum log-likelihood for one changepoint (two regions) over the sequence interval $[1, S]$ as
$$L^2_{1,S} = \max_{1\leqslant q<S-1}\left(L^1_{1,q} + L^1_{q+1,S}\right)$$
where $L^2_{1,S}$ denotes the maximum log-likelihood for two regions over the interval $[1, S]$, and is evaluated for $S = 2, 3, \ldots, J$. Denote the estimated optimal changepoint for two regions over the marker sequence [1,J] (i.e. q maximizing $L^2_{1,J}$) as $\widehat{\eta}_1$.
3. Construct the maximum log-likelihood for r regions over the interval $[1, S]$ using the maximum-likelihood for $r-1$ regions over the intervals $[1, q]$, $q = r-1, r, r+1, \ldots, S-1$ as
$$L^r_{1,S} = \max_{r-1\leqslant q<S-1}\left(L^{r-1}_{1,q} + L^1_{q+1,S}\right),$$
where $L^r_{1,S}$ represents the maximum log likelihood for r regions over the interval $[1, S]$ and is evaluated for $S = r, r+1, \ldots n$. The estimated optimal changepoints (MLEs) for r regions over the binary sequence [1,J] (i.e. the changepoints corresponding to $L^r_{1,n}$) will be denoted by $\widehat{\eta}_1, \widehat{\eta}_2, \ldots, \widehat{\eta}_{r-1}$.
4. Step 3 is repeated for $r = 3, \ldots, R$.

The proof that $L^R_{1,n}$ is the maximum likelihood follows by mathematical induction and by noting that the maximum likelihood for $r-1$ regions over the interval $[1, S]$ must include the maximum likelihood for $r-1$ regions over the interval $[1, j]$ for some $j < S$ (Auger and Lawrence, 1989).

The independence approach assumes that there is no heterogeneity in the binary data. Heterogeneity can be modeled by estimating b_i for each individual in (1). We present the following algorithm to obtain the approximate maximum-likelihood estimates of $\eta_1, \eta_2, \ldots, \eta_{R-1}$:

1B Estimate the individual effects b_i as $\text{logit}(\overline{Y}_{i.})$, where $\overline{Y}_{i.} = \sum_{j=1}^{J} y_{ij}/J$. Denote these estimators as $\widehat{\hat{b}}_i$. Individuals with all 0s or 1s will be dropped from the analyses since they contribute no information about the location of changepoints. The resulting $\widehat{\hat{b}}_i$ will be treated as known in the following steps.

2B Evaluate the log-likelihood for regions $[l, m]$, $l \leqslant m$, with a constant individual frequency, where

$$L^1_{l,m} = \sum_{i=1}^{I} \left[X_{ilm} \log \widehat{p}_{ilm} + (N_{ilm} - X_{ilm}) \log(1 - \widehat{p}_{ilm}) \right]$$

where X_{ilm} and N_{ilm} are the number of positive responses and the total number of binary measurements for the ith subject over the region $[l, m]$. Also note that $\widehat{p}_{ilm}$ is evaluated by estimating $\beta_{l,m}$ in a logistic regression model where $\text{logit} p_{ilm} = \beta_{l,m} + \widehat{\hat{b}}_i$, and where $\widehat{\hat{b}}_i$ is treated as an offset (McCullagh and Nelder, 1989). Thus, $\widehat{p}_{ilm} = \text{logit}^{-1}(\widehat{\beta}_{l,m} + \widehat{\hat{b}}_i)$, where $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$.

We preceed with steps 2–4 as described above.

The above procedure provides us with an approximate maximum-likelihood since the parameters that induce individual heterogeneity are estimated only in step 1B. The full MLE can be obtained with additional computation. For a given number of regions R, we obtain the approximate MLE as described in the steps above. We next fit a logistic regression where we estimate individual effects b_i in the presence of differing regions in the mean structure (estimated from the previous approximate MLE) and then re-compute the approximate MLE treating the estimated b_i as offsets. Specifically, for obtaining the maximum-likelihood for R regions, we fit the model

$$\text{logit} P(Y_{ij} = 1) = \sum_{r=1}^{R} \beta_r I_{(\widehat{\eta}_{r-1}, \widehat{\eta}_r]} + b_i, \tag{2}$$

with $\widehat{\eta}_1, \widehat{\eta}_2, \ldots, \widehat{\eta}_{R-1}$ obtained from step 4, $\widehat{\eta}_1 = 1$, and $\widehat{\eta}_R = J$. We then iterate between steps 2B, 2–4 and expression (2) until we get convergence.

It is well known that maximum-likelihood estimation of the model parameters for (1) will be inconsistent for fixed J as $I \to \infty$ since the number of parameters goes off to infinity (Neyman and Scott, 1948). However, this will not be a practical problem when the number of binary measurements on each individual is large as in our motivating example and in many molecular genetics studies.

Model selection is a difficult and important problem in identifying multiple changepoints. We propose a penalized likelihood approach for selecting the number of changepoints. Specifically, we implement the algorithms for $R = 2, \ldots, J$ and choose the number of changepoints that minimizes the penalized likelihood

$$\text{PLIK}(R) = -2L^R_{1,J} + 2(R - 1)c, \tag{3}$$

where R is the number of regions and c is a penalization constant. Various forms of penalized likelihood have been considered in the literature. Akaike's information criterion (AIC) (Akaike, 1973) fixes $c = 1$. The Bayesian information criterion (Schwartz, 1978) fixes $c = \log(\text{sample size})/2$.

A major inferential goal for our motivating example is to determine whether there are any regions of inflated LOH frequency. Since searching for a tumor suppressor gene is expensive and time consuming, we choose a penalty function so that we would rarely identify one or more changepoints (more than one region) if, in fact, there were none. Thus, we controlled the type I error rate of identifying one or more changepoints when none existed at 5%. The penalization constant c needs to be computed separately under the independence, approximate MLE, and full MLE approaches. For the independence approach, we estimate c with the following steps:

(i) We estimate the model parameters with no changepoints from the data. This is done assuming a constant frequency over all the binary sequences. Specifically, $\widehat{p} = X_{1J}/N_{1J}$.
(ii) We simulate 10 000 datasets where sequences of independent Bernoulli data are generated with proportion $\widehat{p}$. We implement steps 1–4 for $R = 1, 2, 3, \ldots, J$. The penalization constant c in (3) is calculated such that 95% of the time PLIK(R) achieves its minimum at $R = 1$.

For the approximate and full MLE approach allowing for patient heterogeneity, we estimate c as follows:

(i′) We estimate the model parameters with a constant frequency for each individual, $\widehat{p}_i = X_{i1J}/N_{i1J}$.
(ii′) We simulate 10 000 datasets where sequences of independent Bernoulli data, for each patient, are generated with proportion $\widehat{p}_i$. We implement the approximate or full MLE as previously described. The penalization constant c is chosen as described in (ii) above.

As we will see in Section 4, in most practical situations where the change in frequency between regions (at the changepoints) is large, this penalty function will perform well in estimating the correct number of changepoints.

The parametric bootstrap (Efron and Tibshirani, 1993) was used to assess the variability in changepoint estimation and to estimate confidence intervals for the marginal means, which can be displayed graphically. Bootstrap samples were obtained by generating data from model (1) with parameter estimates (including number and location of changepoints) and missing data pattern obtained from the original dataset (10 000 bootstrap samples). For each bootstrap sample, parameter estimation and model selection was done as described earlier. Ideally, we would need to re-estimate the c (using (i′) and (ii′)) for model selection in every bootstrap sample. This is not computationally feasible. One advantage of the parametric as compared to the nonparametric bootstrap is that c is very stable over replicate samples. Therefore, we use the c estimated from the actual data to choose models in all bootstrap samples. Pointwise 95% confidence intervals around estimated marginal means, obtained from the bootstrap, are used to assess variability. In addition, other measures of variability are presented, such as the proportion of times (bootstrap samples) a particular marker is included in the region with the highest LOH frequency.

3. Analysis

We fit the full MLE, approximate MLE and the independence approaches to the LOH data on chromosome 13. Table 1 shows the estimated changepoints (for up to five changepoints) and the associated penalized likelihood for models fit to the data for each of the three approaches. For each approach, the optimal number of changepoints corresponds to the model with the smallest penalized likelihood. Based on the simulation methods described in Section 2, the penalization constant was chosen as $c = 5.01, 5.01$, and 4.84 for the three approaches, respectively. These penalization constants are larger than those prescribed by AIC ($c = 1$) or BIC ($c = \log(56 \times 85)/2 = 4.2$). The approximate MLE and the full MLE resulted in identical changepoint estimates for each of the models. The penalized likelihood values were similar for the two approaches. Although the independence model with the smallest

Table 1. *Estimated changepoints and penalized likelihoods for chromosome 13 under the A. full MLE, B. approximate MLE, and C. independence approaches*

	# of changepoints	Estimated changepoints[1]					Penalized likelihood	
A.	0						2417.8	
	1	18.77					2386.1	
	2	16.38	18.08				2358.9	
	3	16.38	18.77	21.00			2350.8	
	4	16.38	18.08	19.25	21.00		2333.1	←
	5	16.38	18.08	19.25	21.00	27.23	2334.5	
B.	0						2417.8	
	1	18.77					2386.7	
	2	16.38	18.08				2360.2	
	3	16.38	18.77	21.00			2353.6	
	4	16.38	18.08	19.25	21.00		2336.4	←
	5	16.38	18.08	19.25	21.00	27.23	2338.2	
C.	0						3595.1	
	1	18.08					3582.3	
	2	16.38	18.08				3564.6	
	3	16.38	18.08	21.00			3567.5	
	4	16.38	18.08	19.25	21.00		3557.9	←
	5	16.38	18.08	19.25	21.00	21.76	3561.8	

[1] Distance in Mb from the p-telomere

penalized likelihood (four changepoints) resulted in the same changepoints as the methods that accounted for heterogeneity, changepoint estimates were different for other numbers of changepoints. A comparison of the three approaches on the LOH data leads one to question (i) whether the additional computational intensity of the full MLE approach is worth the effort, and (ii) whether the independence model does poorly for heterogeneous data.

Figure 3 shows the estimated changepoints and marginal LOH frequencies on chromosome 13 for the 56 esophageal cancer patients. Parametric bootstrap pointwise 95% confidence intervals demonstrate that the marginal frequencies can be estimated with reasonable precision. We identify an inflated region that included four markers at locations 19.37, 19.49, 20.49, and 21.00 Mb. This inflated region is the same region identified by Li *et al.* (2001) using their ad hoc method. Figure 4 shows the proportion of bootstrap realizations (10 000 bootstrap samples were performed) in which a specific marker is located within the region with the highest LOH frequency. Markers at 19.37, 19.49, 20.49, and 21.00 Mb are contained in this inflated region in 96, 99, 99, and 97% of the bootstrap samples, respectively. Thus, this inflated region is precisely estimated. As mentioned in the Introduction, there are numerous genes in this region which are thought to play a role in esophageal cancer. Future laboratory work will focus on identifying whether any of these genes are tumor suppressor genes.

4. SIMULATIONS

We conducted a series of simulation studies to examine the properties of the various approaches. Our first simulation corresponded to the LOH example ($I = 56$ and $J = 85$). We simulated data with changepoints at locations 16.38, 18.08, 19.25, and 21.00 Mb (see Figure 3) with parameters (β_r and b_i) chosen as the ones estimated using the full MLE approach on our LOH data. Thus, there were large

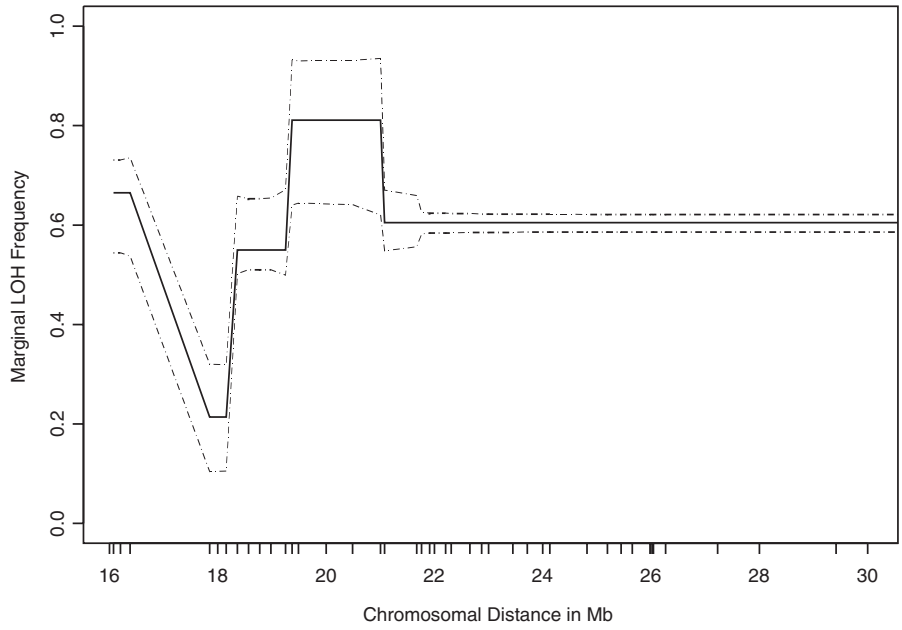

Fig. 3. Approximate maximum-likelihood estimates of the marginal LOH frequency (lines) and changepoints for markers on chromosome 13. Pointwise 95% confidence intervals are presented. Estimates and intervals are only presented up to a distance of 30 Mb from the p-telomere. Estimates and intervals are constant for all markers greater than 21 Mb from the p-telomere.

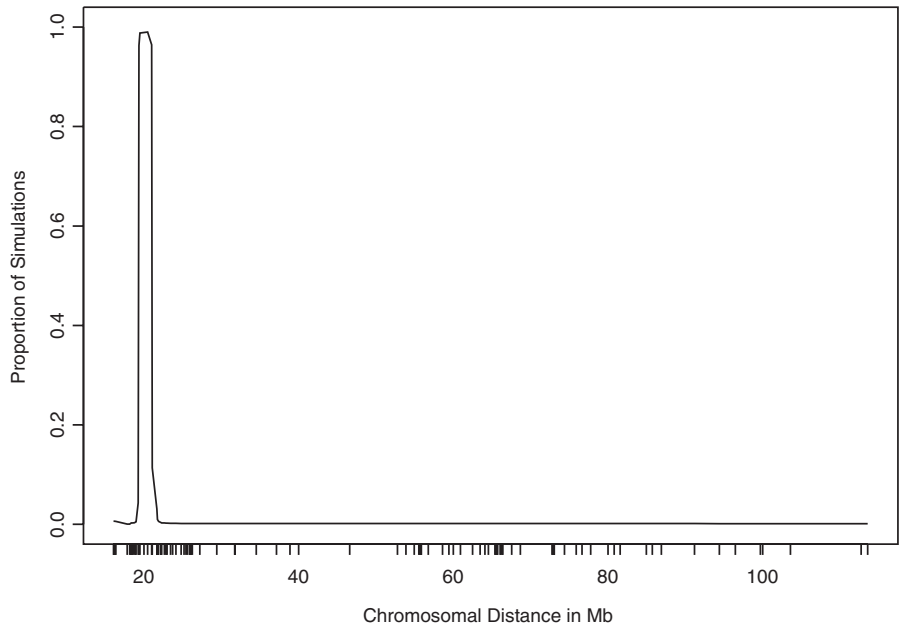

Fig. 4. The proportion of bootstrap realizations (10 000 bootstrap samples were performed) in which a specific marker is located within the estimated region with the highest LOH frequency.

Table 2. *Simulation in which data were generated corresponding to the four changepoint model for chromosome 13. Changepoints and marginal frequencies are given in Table 1 and in Figure 3. Changepoints are generated at a distance of 16.38, 18.08, 19.25, and 21.00 Mb from the p-telomere. Penalization constants, c, were calculated separately for each approach as described in Section 2. 10 000 simulated datasets with the same data structure ($I = 56$ and $J = 85$) and missing data pattern were generated*

	Frequency of models selected %		
# of changepoints	Full m.l.e	Approximate MLE	Independence
0	0	0	0
1	0	0	0
2	0	1	14
3	0	0	3
4	95	96	83
5	4	3	0
6	0	0	0

differences in the frequency across regions and substantial patient heterogeneity. Data were generated with the same frequency of missingness (due to uninformative markers) as in the example. The simulation results were based on 10 000 simulated datasets. The penalization constants for the full, approximate, and independence approaches were chosen as the values estimated for the example (Section 3). Table 2 shows the results of this simulation. The full and approximate MLE approaches detected the correct number of changepoints 95 and 96% of the time. In fact, the results for the approximate and full MLE approaches were very similar across realizations; 99% of the simulated datasets resulted in a final model with the same number and location of changepoints for the two approaches. The changepoints at locations 16.38, 18.08, 19.25, and 21.00 Mb were detected 99, 99, 91, and 84% for both the full MLE and the approximate MLE approaches. Additional simulations, with varying sample sizes (I and J) and number of changepoints also showed that the full and approximate MLE had very similar operating characteristics (data not shown). Thus, there is no benefit for the additional computational intensity of the full MLE approach. These simulations also showed that edge effects can be important. Specifically, we saw that it is more difficult to identify a changepoint near the edge of a chromosome than in the middle.

Our approach for choosing the penalization constant in (3) appeared to work well. We consistently saw that, for large differences in frequency between regions, the correct number of changepoints was identified with high probability using this approach (see Table 2). For the simulation presented in Table 2, the penalization constants were estimated to be $c = 5.01$ for the full and approximate MLE approaches, respectively. This penalization is larger than either the AIC or BIC penalized likelihood approaches ($c = 1$ for AIC and $c = \log(56 \times 85)/2 = 4.2$ for BIC). Thus, one would expect that AIC and BIC would have a tendency to choose models with too many changepoints. Additional simulations as described in Table 2 (four changepoints) were performed to verify this. The AIC penalization chose models with four and five changepoints in 3% and 97% of the simulated datasets. The BIC penalization chose models with four and five changepoints in 92% and 8% of the realizations. In either case, our procedure outperformed these traditional penalized likelihood approaches.

The independence approach did not perform as well as the approximate or full MLE (Table 2). There was a tendency for the independence approach to choose models with too few changepoints; a model

with the correct number of changepoints was only chosen in 83% of the realizations. The changepoints at locations 16.38, 18.08, 19.25, and 21.00 Mb were detected in 98%, 98%, 81%, and 77% of the simulated datasets with the independence approach. With additional simulations (again with different sample sizes and number of changepoints), we consistently showed the poorer performance of the independence model (data not shown).

Simulations were performed to examine the sensitivity of inferences to important modeling assumptions. We examined robustness to the assumption of spatial independence by generating data according to the model

$$\text{logit}P(Y_{is(j)} = 1) = \sum_{r=1}^{R} \beta_r I_{(\eta_{r-1},\eta_r]} + b_i + \epsilon_{i,s(j)}, \tag{4}$$

where $s(j)$ is the chromosomal distance in Mb from the p-telomere for the jth marker and $\epsilon_{i,s(j)}$ is a Gaussian process called the Ornstein–Uhlenbeck process (Karlin and Taylor, 1981) which has mean zero and spatial covariance structure given by

$$\text{Cov}(\epsilon_{i,s(j)}, \epsilon_{i,s(j')}) = \theta_1 \exp(-\theta_2 |s(j) - s(j')|).$$

We simulated data according to (4) with the number and location of changepoints, β_r, and b_i estimated from the example data and with the same missing data pattern as the example. We chose $\theta_1 = 1$ and $\theta_2 = 0.02$, which induced a sizable spatial dependence which diminishes slowly with increasing chromosomal distance. With these parameters the latent random variables $\epsilon_{i,s(j)}$ have correlation 0.82, 0.67, and 0.45 when they are separated by 10, 20, and 40 Mb, respectively. This is substantially higher spatial correlation than we observed in our LOH dataset. The penalization constant was chosen by simulating 10 000 datasets according to (4) with parameters chosen as those estimated from the LOH data. The constant c was chosen by estimating each p_i by averaging the simulated sequence for that individual and then proceeding with step (ii′) as described in Section 3. The approximate MLE approach (which ignores spatial dependence) chose the correct number of changepoints in 70% of the 10 000 simulated datasets. There was a tendency to choose the optimal number of changepoints too low; for example, only two changepoints were identified in 25% of the simulated datasets. The changepoints at locations 16.38, 18.08, 19.25, and 21.00 Mb were detected in 91, 88, 67, and 62% of the simulated datasets. Figure 5 shows the proportion of simulated realizations in which a specific marker is located within the estimated region with the highest LOH frequency. Even with sizable spatial correlation, the true high region located between 19.37 and 21.00 Mb is included within the identified high region in most of the simulation realizations. Markers at 19.37, 19.49, 20.49, and 21.00 Mb are included within the identified high region in 86%, 88%, 88%, and 86% of the simulation samples.

We examined the robustness of inferences to the possibility that LOH changes according to a smooth rather than a changepoint process. We fit a model with a separate marker and individual effects,

$$\text{logit}P(Y_{ij} = 1) = \beta_j + b_i, \tag{5}$$

to our LOH data. We simulated datasets according to (5) with parameter values estimated from the example dataset and with missing data pattern as in the example. The resulting marginal mean for this model is shown in Figure 6. As in the observed data we see a large ‘peak’ between markers 19.25 and 21.08 Mb and a smaller increase between 31.74 and 46.61 Mb (these correspond to the two regions identified by Li *et al.* (2001) as deletion regions). Interest focuses on whether we are able to identify these regions. Figure 7 shows the proportion of simulated datasets in which a marker is included within the estimated region with the highest LOH frequency. The figure demonstrates that one of these two regions is most often chosen as the region with highest LOH frequency. Specifically, markers at locations 19.37,

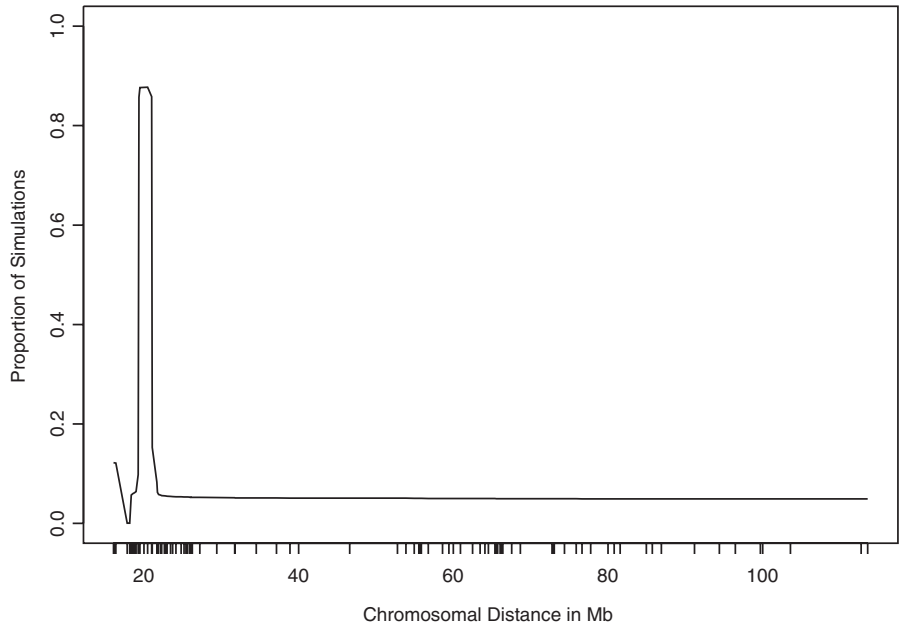

Fig. 5. The proportion of simulations in which a specific marker is located within the estimated region with the highest LOH frequency for data generated with sizeable spatial dependence.

19.49, 20.49, and 21.00 Mb are in this highest region in 44%, 53%, 56%, and 56% of the realizations, and markers at locations 31.74, 31.79, 34.52, 37.14, 38.88, and 46.61 Mb are in the highest region in 27%, 28%, 28%, and 26%, 22%, and 12% of the realizations. Thus, our procedure works well in identifying regions of high frequency even if the data do not truly follow a changepoint process.

5. DISCUSSION

We developed a procedure for identifying changepoints for heterogenous binary data which was motivated by our scientific interest in identifying a tumor suppressor gene for esophageal cancer. Our scientific interest was on localizing regions with high LOH frequency so that candidate tumor suppressor genes could be identified. This was particularly challenging for the LOH data since we had a high overall LOH frequency and a large between-subject variation in this frequency. Using our methodology, we localized an inflated region which included four LOH markers; this region contains a set of candidate genes that will be studied further in future studies using more refined genetic procedures (e.g. mutational analysis).

We proposed an approximate MLE approach which accounted for heterogeneity and compared the performance of this procedure to an independence model and a full MLE approach. The approximate MLE approach had improved performance relative to the independence model, and there was no apparent advantage in applying the more computationally intensive full MLE approach. The model accounts for heterogeneity in LOH frequency through estimating individual-specific intercepts. Although the approach is appealing since it makes no distributional assumptions on the individual effects b_i, it results in inconsistent estimation when J is small (this was not an issue in our LOH example since $J = 85$). An extension of conditional logistic regression (Agresti, 1990) for the multiple changepoint problem would be

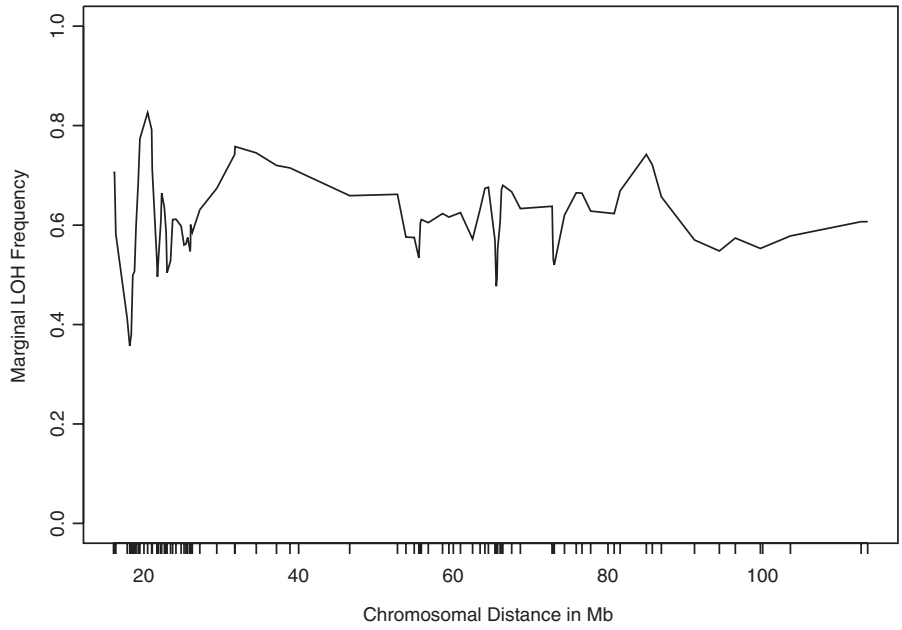

Fig. 6. Marginal means estimated from model (5).

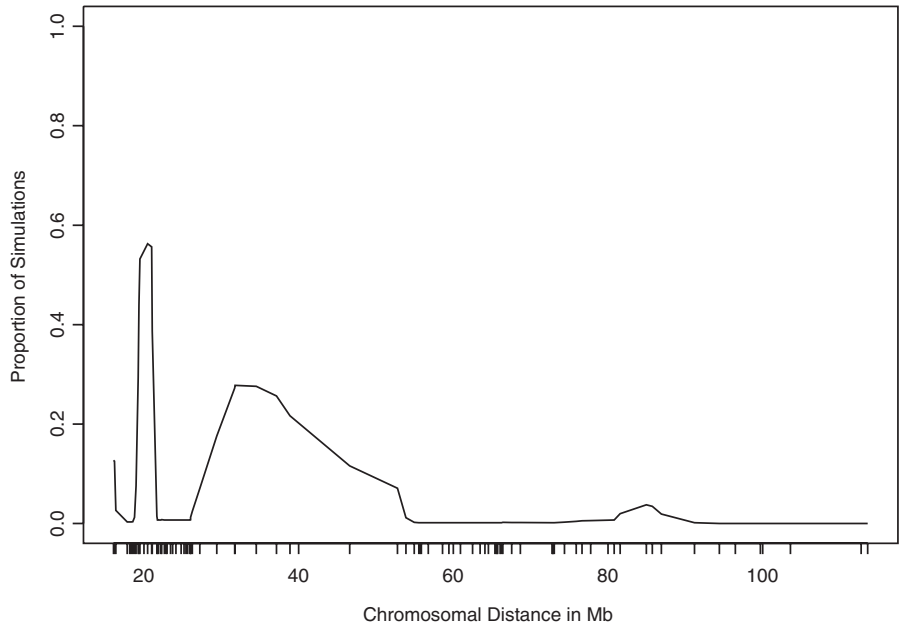

Fig. 7. The proportion of simulations in which a specific marker is located within the estimated region with the highest LOH frequency for data generated with a smooth curve and analyzed with a changepoint model.

appropriate for small J; this is an area of future research. Our approach treats heterogeneity as a nuisance and, therefore, would not be useful for making inference about the heterogeneity across individuals. A binary random effects model (i.e. generalized linear mixed model) formulation may be more appropriate when estimating the heterogeneity is of interest. The development of such a random effects model with multiple changepoints is an area for future research.

The model assumed that there is no spatial correlation in the binary sequences. This was a reasonable assumption for the LOH data since spatial dependence was negligible (see Figure 2). In addition, we showed using simulation that our methodology worked well in identifying inflated regions even with sizeable spatial correlation. However, when spatial dependence exists, an approach which incorporates this spatial correlation would be expected to have increased performance. Developing multiple changepoint approaches for spatially correlated data is an area of future research. Our formulation also assumed that LOH frequency follows a changepoint structure. Using simulations, we showed that the highest regions are often identified even when the changes in LOH frequency were more gradual than a multiple changepoint structure would allow.

Our scientific strategy was to localize regions with inflated LOH frequency, identify candidate genes within those regions, and use other genetic techniques to test whether these candidate genes are, in fact, tumor suppressor genes. Since searching for tumor suppressor genes is expensive and time consuming, we chose a penalty function so that we would rarely identify a changepoint if, in fact, there were none. Specifically, we estimated the penalized constant through simulation by controlling the type I error rate at 5% of falsely concluding that there were more than one region (one or multiple changepoints) when in fact there was only one region (no changepoints). Our simulation studies showed that, in most practical situations in which the change in frequency between regions (at the changepoints) was large, this penalty performed well in estimating the correct number of regions. Our procedure did better than AIC or BIC, which penalize too little, and generally resulted in choosing models with too many regions. An alternative to our penalization procedure is a Bayesian approach in which a prior distribution for the number of changepoints is chosen so that 95% prior probability is placed on the zero-changepoint model. A Bayesian approach to our problem is an area of future research.

We had other scientific interests in this study including whether the pattern in LOH frequency varies across family history status (i.e. did a first or second degree relative have gastrointestinal cancer?). Such a comparison may help distinguish whether important genetic mutations are somatic or germline. Developing methods for comparing the processes between groups is an area of future research.

ACKNOWLEDGEMENTS

The authors thank Professor Scott Zeger, an Associate Editor, and two reviewers for helpful comments on the original version of the manuscript. We thank the Center for Information Technology, National Institutes of Health, for providing access to the high performance computational capabilities of the Beowulf cluster computer system.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. New York: Wiley.

AKAIKE, H. (1973). Information theory and an extension of the maximum-likelihood principle. In Petrov, B. N. and Csaki, F. (eds), *Second International Symposium on Information Theory*, Budapest: Akademia Kaido, pp. 267–281.

AUGER, I. E. AND LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology* **51**, 39–54.

BRAUN, J. V. AND MULLER, H. G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science* **13**, 142–162.

BRAUN, J. V., BRAUN, R. K. AND MULLER, H. G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87**, 301–314.

CARLIN, B. P., GELFAND, A. E. AND SMITH, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics* **41**, 389–405.

CHURCHILL, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

EFRON, B. AND TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

FU, Y. X. AND CURNOW, R. N. (1990). Maximum-likelihood estimation of multiple changepoints. *Biometrika* **77**, 563–573.

GRUICE, N. A., ABALONE, E. C. A., BARDOEL, A. F. J., DEVILEE, P., FRANTS, R. R. AND CORNELISSE, C. J. (1993). PRC-based microsatellite polymprphisms in the detection of loss of heterozygosity in fresh and archival tumor tissue. *British Journal of Cancer* **64**, 308–313.

HU, N., ROTH, M. POLYMEROPOLOUS, M. *et al.* (2000). Identification of novel regions of allelic loss from a genomewide scan of esophageal squamous-cell carcinoma in a high-risk Chinese population. *Genes, Chromosomes and Cancer* **27**, 217–228.

HUANG, J., HU, N. GOLDSTEIN, A. M. *et al.* (2000). High frequency allelic loss on chromosome 17p13.3-p11.1 in esophageal squamous cell carcinoma from a high incidence area in northern china. *Carcinogenesis* **21**, 2019–2026.

LI, G., HU, N. GOLDSTEIN, A. M. *et al.* (2001). Allelic loss on Chromosome Bands 13q11-q13 in esophageal squamous cell carcinoma. *Genes, Chromosomes and Cancer* **31**, 390–397.

KARLIN, S. AND TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. New York: Academic Press.

KORN, E. L. AND WHITTEMORE, A. S. (1979). Methods for analyzing panel studies of acute health of air pollution. *Biometrics* **35**, 795–802.

MCCULLAGH, P, AND NELDER, J. A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

NEWTON, M. A. AND LEE, Y. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**, 1088–1097.

NEYMAN, J. AND SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

ROSS, D. W. (1998). *Introduction to Oncogenes and Molecular Cancer Medicine*. New York: Springer.

SCHWARTZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

STEPHENS, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics* **43**, 159–178.

[*Received 14 March 2003; first revision 29 October 2003; second revision 10 March 2004; accepted for publication 11 March 2004*]